Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media

Aye or naw, whit dae ye hink?
Scottish independence and linguistic identity on social media

Philippa Shoemark[*]
p.j.shoemark@ed.ac.uk

Debnil Sur[†]
debnil@stanford.edu

Luke Shrimpton[*]
luke.shrimpton@ed.ac.uk

Iain Murray[*]
i.murray@ed.ac.uk

Sharon Goldwater[*]
sgwater@inf.ed.ac.uk

[*]School of Informatics
University of Edinburgh

[†]Department of Computer Science
Stanford University

Abstract

Political surveys have indicated a relationship between a sense of Scottish identity and voting decisions in the 2014 Scottish Independence Referendum. Identity is often reflected in language use, suggesting the intuitive hypothesis that individuals who support Scottish independence are more likely to use distinctively Scottish words than those who oppose it. In the first large-scale study of sociolinguistic variation on social media in the UK, we identify distinctively Scottish terms in a data-driven way, and find that these terms are indeed used at a higher rate by users of pro-independence hashtags than by users of anti-independence hashtags. However, we also find that in general people are *less* likely to use distinctively Scottish words in tweets with referendum-related hashtags than in their general Twitter activity. We attribute this difference to style-shifting relative to audience, aligning with previous work showing that Twitter users tend to use fewer local variants when addressing a broader audience.

1 Introduction

A central idea from sociolinguistics is that people's social identity is reflected in their use of language, and that people modulate their use of language in order to present particular identities in different situations. The recent availability of social media data has raised interest in confirming and extending these results using large scale datasets. For example, Twitter data has been used to examine patterns of regional variation in general US English (Doyle, 2014; Huang et al., 2015), African American English (Jones, 2015), and global Spanish (Gonçalves and Sánchez, 2014), and to study variation associated with factors such as race/ethnicity (Jones, 2015; Blodgett et al., 2016; Jørgensen et al., 2015) and gender (Bamman et al., 2014). These studies have shown that tweets mirror spoken language in many ways, such as displaying dialect variation not only in the use of distinct lexical items, but also in the use of non-standard spellings to indicate non-standard pronunciation—in fact, these spellings even reflect the phonological processes found in spoken language (Eisenstein, 2015). There is also evidence that, as in spoken language, individuals may shift their style of language in response to the audience. In particular, studies have found that when the expected audience of a tweet is larger, Americans use fewer non-standard and local words (Pavalanathan and Eisenstein, 2015) and Dutch bilingual speakers of a minority language are more likely to use Dutch rather than their other language (Nguyen et al., 2015). A small-scale case study of a single Scottish Twitter user also provides preliminary evidence that users may modulate their production of regional variants according to the topic of the tweet (Tatman, 2015).

Here we present the first large-scale sociolinguistic study of British tweets, and the first to examine the relationship between sociolinguistic variation and political views using social media data. We use a large corpus of tweets to examine the relationship between users' linguistic choices and their views about the 2014 Scottish independence referendum. The referendum (on whether Scotland should leave the UK) generated considerable political discussion and an unprecedented turnout of 84.6% of the

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1239–1248,
Valencia, Spain, April 3-7, 2017.

electorate, with the 'No' (anti-independence) side taking 55.3% of the vote. The 2013 Scottish Social Attitudes Survey (ScotCen, 2013) showed a clear correlation between national identity and voting intentions (53% of those who identified as 'Scottish not British' said they intended to vote 'Yes' to independence, vs. just 5% of those who identified as 'British not Scottish'), and there was much discussion in the popular press about the relationship between a sense of Scottish identity and support for Scottish sovereignty.

Although this recent discussion was not centered on language, there is a long history of scholarly discourse connecting the use of the Scots language[1] and sociolinguistic and political identity (Grant, 1931; Mcafee, 1985; Corbett et al., 2003). If this connection still holds today, then we might expect to find that those on the 'Yes' side of the debate use more identifiably Scottish language than those on the 'No' side. We might also expect to find some modulation of Scottish language use depending on whether users are discussing the referendum or not.

To examine these questions, we used a data-driven approach to identify linguistic terms that are used more in Scotland than in the rest of the UK. The identified terms include uniquely Scots words that are attested in Scots literature dating back to the 1600s and earlier, contemporary regional colloquialisms, spelling variants of Standard English words which reflect Scottish pronunciations, and acronyms used as shorthand for distinctive Scottish phrases. From these, we selected variables for which users can produce either a Standard English or Scottish variant (e.g., DO vs. DAE). We then classified users as pro- or anti-independence based on the referendum-related hashtags they used and asked whether these two groups use Scottish variants at different rates. We found that the pro-independence group did use Scottish variants significantly more than the anti-independence group, although the overall rate of Scottish variants is very low amongst all users.

Next, we compared the use of Scottish variants in tweets containing referendum-related hashtags to their use in other tweets. If users are aiming to project their Scottish identity as part of political discourse, then we might expect greater use of Scottish variants in referendum tweets than in non-referendum tweets. However, previous studies have suggested that non-standard and local variants are used *less* frequently in tweets containing hashtags, which typically have a larger audience than other tweets (Pavalanathan and Eisenstein, 2015). This effect would predict the opposite result—a lower use of Scottish variants in tweets with referendum hashtags—and indeed this is the result we found. So it appears that although pro-independence users do make greater use of Scottish variants overall, they do not increase their Scottish usage when engaging in broad-audience political discourse.

To summarize, the contributions of our paper are: (1) The first large-scale study of dialect variation on twitter in the UK. We show that in addition to using Scots in speech and some literary genres such as poetry, people are using Scots in informal public writing. The data-driven approach enables us to identify Scotland-specific lexical items without relying on pre-conceived notions of which variables to look for (cf. Tatman, 2015), and reveals that in addition to using attested Scots vocabulary, Twitter users appear to be creatively adapting to the medium with their use of acronyms for distinctly Scottish turns of phrase. (2) The first study connecting sociolinguistic variables to political stance using social media data, showing that pro-independence users have a higher rate of Scottish usage. (3) Further evidence of Pavalanathan and Eisenstein's (2015) claim that Twitter users modulate their language according to the audience, with local variants being less likely in tweets directed to larger audiences.

2 Context

'Scots' refers to the group of dialects historically spoken in the Lowlands of Scotland. While Scots has Anglo-Scandinavian origins in common with English, by the 16th century its pronunciation, vocabulary, and literary norms had considerably diverged from those of English, and Scots had become established as the prestige language in Scotland (Kay, 1988).[2] However, following the Union of Crowns in 1603, when King James VI of Scotland acceded to the thrones of England and Ireland,

[1] Historically, Scots has been considered a different language than English (see §2), though with many cognates and overlapping vocabulary. Most native Scottish people today speak some variety of Scottish English, which retains a few uniquely Scots words but is mainly distinguished from other varieties of English by its pronunciation.

[2] Previously, Gaelic had been the dominant spoken and literary language in Scotland. Note that while in medieval times non-Gaelic speakers referred to the Gaels as 'Scots', what we now refer to as 'Scots' is the Anglo-Scandinavian language which spread at the expense of Scottish Gaelic (a Celtic language) in the 15th & 16th centuries.

he and his court began to adopt English norms in their writing. After the Union of Parliaments in 1707, English firmly replaced Scots as the language of serious or elevated discourse in Scotland (Grant, 1931). While some people still use distinctive elements of Scots in their speech, until recently the average Scottish person's exposure to written Scots would have been largely confined to a select few literary domains such as poetry and comic narrative (Corbett et al., 2003). However, social media has given rise to a new genre of casual, communicative writing that is potentially visible to large and diverse audiences, providing both a platform and an impetus to express one's identity through the use of written language. Below, we provide three example tweets (each from a different user) which contain orthographic representations of Scots vocabulary and/or Scottish English pronunciation. Standard English variants of Scottish terms are provided in italics.

(1) No matter how shite *[shit]* a day you've had just remember there's always good biscuits in yer *[your]* grannies hoose *[house]*

(2) "Absolute carnage" at polling station earlier. Bairns *[kids]* playing, polite grannies, Yessers and Nos blethering *[blathering]* to each other. #VoteYesScotland

(3) #fuckoffscotland hud on we will fuck off but afore we dae eh challenge ye tae a square go ya queen loving DIDDY doughnut Sasijs YUP-TAE
#fuckoffscotland hold on we will fuck off but before we do I challenge you to a fair fight you queen loving fools. What are you doing!?

3 Data

Our data was drawn from the Sample endpoint of Twitter's Streaming API (a.k.a. the 'Spritzer'), which provides a random 1% sample of all public tweets in near real-time. We started with all tweets streamed from the Spritzer between 1st September 2013 and 30th September 2014. These dates cover a year of activity leading up to the referendum, as well as the day the vote took place (18 September 2014), and immediate reactions. We used a language classifier (Lui and Baldwin, 2012) to filter out non-English tweets, yielding an initial dataset of 629,431,509 tweets.[3] Because we are interested in the linguistic choices that individuals make in various contexts, we took steps to remove tweets which were not originally authored by the individual who posted them. Retweets (tweets which are verbatim copies of other tweets) were identified by a case-insensitive search for the token 'RT', and discarded. Quote tweets (tweets which contain verbatim copies of other tweets, but are augmented with original comments) were dealt with by discarding any text between double quotation marks, but retaining the remainder of the tweet.

From this initial dataset we extracted three overlapping subsets:

The Geotagged-UK (GU) dataset contains all tweets geotagged to a location in the United Kingdom (1,654,204 tweets by 446,923 distinct users).

The Geotagged-Scotland (GS) dataset contains all tweets geotagged to a location in Scotland (166,992 tweets by 40,861 distinct users).

The Indyref Tweets (IT) dataset consists of tweets containing hashtags relating to the 2014 Scottish Independence Referendum.

To construct the IT dataset, we first created a list of relevant hashtags, starting with the following five seed hashtags: *#IndyRef, #VoteYes, #VoteNo, #YesScotland, #BetterTogether*.[4] For each of these five seeds, we extracted from our initial filtered dataset a list of *all* tweets by any user who used the seed hashtag. We identified the 100 most frequent hashtags in each of these five lists of tweets, and manually discarded all hashtags which were unrelated to the referendum, as well as those which were highly ambiguous (e.g., *#Indy*, which sometimes refers to the referendum, but also commonly refers to a genre of music). The resulting list of referendum-related hashtags is given in Table 1.

Next, we extracted all tweets from our initial dataset which contain at least one of the hashtags on this list, yielding 77,708 tweets by 26,019 distinct users. We then applied a heuristic to filter out tweets produced by bots and spammers: for

[3] One might be concerned that an automatic language filter could remove some of the heavily Scottish tweets. However, even tweets such as example (3) in §2 are assigned a very high probability of being English by the filter. Perhaps other tweets with many Scottish terms were filtered out, in which case we will underestimate the probability that users choose Scottish variants. However this issue should not cause us to find differences in use between different groups where there are none.

[4] 'Yes Scotland' and 'Better Together' are the names of the principal organisations representing the Yes and No vote campaigns, respectively.

each user in the IT dataset for whom we had at least 5 tweets in the initial dataset, we computed the proportion of their tweets that contain URLs, and discarded users for whom this proportion was in the 90th percentile. This step filtered out 11,443 tweets by 1389 users.

Note that seven of the hashtags in Table 1 (*#voteyes, #bettertogether, #nothanks, #voteno, #yes2014, #letsstaytogether,* and *#yesvote*) are occasionally used in contexts unrelated to the Scottish Independence Referendum (e.g. *#bettertogether* can also refer to interpersonal relationships). However, they are distinctive enough that if a user has also used hashtags which are unambiguously related to the referendum, then it seems reasonable to assume that their usage of these potentially-ambiguous hashtags relates to the referendum too. Therefore, in order for a tweet containing one of these seven hashtags to be retained in the Indyref dataset, we required that its author had also used at least one other hashtag from Table 1. This criterion filtered out a further 6601 tweets by 6041 distinct users, such that the final IT dataset contains 59,664 tweets by 18,589 distinct users.

4 Identifying distinctively Scottish vocabulary on Twitter

We wish to identify terms that are more likely to be used by Twitter users in Scotland than in the rest of the UK. We follow the method of Pavalanathan and Eisenstein (2015), who used the Sparse Additive Generative Model of Text (SAGE) framework (Eisenstein et al., 2011) to identify tweet terms associated with metropolitan areas in the United States. SAGE models deviations in the log-frequencies of terms in a corpus of interest (here, the GS dataset) with respect to their log-frequencies in some "background" corpus (here, the GU dataset). The estimated deviations are regularized to avoid overstating the importance of deviations in the frequencies of rare words. Here, we use a publicly available implementation of SAGE[5] to obtain log-frequency deviation estimates for all terms which occur at least fifty times in the GU dataset, excluding hashtags, mentions, URLs, and stopwords. The terms with the highest estimates are those which are most distinctive to tweets geo-located in Scotland.

[5]https://github.com/jacobeisenstein/jos-gender-2014/

4.1 Scotland-specific terms

Unsurprisingly, many of the Scotland-specific terms are proper nouns which are topically associated with Scotland, such as Scottish placenames, political figures, and sports personalities. There are also several common nouns (e.g. 'devolution', 'bagpipes') and verbs (e.g. 'canvass', 'invade') which are strongly associated with the political or cultural climate in Scotland. These terms occur with greater relative frequency in the GS dataset simply because their referents are discussed with greater relative frequency; not because they are distinct from the terms that people in the rest of the UK use to index those referents. However, there are also many terms with high log-frequency deviations that *are* linguistically distinctive. To isolate such terms, we began with the 400 terms with the highest estimated deviations, and then manually filtered this list, discarding Standard English words, proper nouns, numerals, and non-standard terms which had clear topical associations (e.g. 'devo': an abbreviation for 'devolution'; 'hh': an acronym for 'Hail Hail', a football chant used by supporters of Celtic F.C.). The remaining 113 distinctively Scottish terms are listed in Table 2.

Almost three fourths of these terms are attested in the Scottish National Dictionary (SND) (Grant and Murison, 1931) or its online supplement (Scottish Language Dictionaries, 2004), which catalogue words that are distinctive to Scots (i.e. those which are not used, or are used differently, in Standard English), covering the period from the 1700s up to the present day. Many are also attested in the Dictionary of the Older Scottish Tongue (Aitken et al., 1990), which catalogues the entire vocabulary of Scots from the 1100s to the late 1600s. Of the attested Scots words, some are unique to Scots, e.g. BAIRNS ('sons/daughters'), GREETIN ('weeping'); some are cognates with English words that have fallen out of common usage, e.g. CRABBIT ('crabbed'; 'ill-tempered'), FEART ('feared'; 'frightened/timid'); some are cognates with English words but have a wider range of senses, e.g. HUNNERS is cognate with'hundreds', but used more generally to mean 'lots' as in "love you hunners", "there was hunners to do"; and many differ only in form from their English cognates, e.g. AFF ('off') and BAW ('ball').

Of the 29 terms that are not attested in SND, 9 are spelling variants or derived forms of attested

Neutral hashtags: *#IndyRef* (46,491) *#ScotlandDecides* (2552) *#BBCIndyref* (1591) *#ScotDecides* (934) *#BigBigDebate* (676) *#ScottishIndependence* (583) *#IndyPlan* (296) *#ScottishReferendum* (239) *#IndyReasons* (180) *#IndependentScotland* (26)
Yes hashtags: *#VoteYes* (8463) *#YesScotland* (1453) *#YesBecause* (1312) *#The45* (908) *#YouYesYet* (827) *#YesScot* (670) *#ActiveYes* (508) *#HopeOverFear* (325) *#Yes2014* (321) *#VoteYesScotland* (256) *#GoForItScotland* (153) *#The45Plus* (138) *#YesFlash* (114) *#GenYes* (92) *#YesVote* (76) *#1Year2Yes* (56) *#VoteAye* (53) *#FreeScotland* (52) *#SaorAlba* (45) *#YesGenerations* (39) *#RIPBetterTogether* (36) *#NHSForYes* (24) *#AnotherScotlandIsPossible* (23) *#EndLondonRule* (13)
No hashtags: *#BetterTogether* (2342) *#NoThanks* (1103) *#VoteNo* (867) *#LabourNo* (333) *#LetsStayTogether* (145) *#VoteNo2014* (92) *#UKOK* (86) *#VoteNoScotland* (45) *#JustSayNaw* (43) *#VoteNaw* (42) *#NoScotland* (34) *#DayOfUnity* (30) *#MaintainTheUnion* (9)

Table 1: Hashtags related to the Scottish Independence Referendum and their frequencies in the IT dataset

Scots words, e.g. CANA, CANNY, and CANI are alternative spellings of the attested CANNAE, and WANTY is a contracted form of 'want to', analogous to the attested GONNAE and GONY. A further 5 are orthographic representations of distinctively Scottish pronunciations, e.g. ANO ('I know'), HING ('thing'); and 2 are acronyms for distinctively Scottish turns of phrase: GTF ('Get Tae Fuck') and MWI ('Mad Wae It'). The final 13 could be described as contemporary Scottish slang, and include abbreviations: BEVY ('beverage')[6], DEFOS ('definitely'); drug-related lexis: WHITEY, ECCIES; profanities: BOABY, FANNYS; and everyday affective and descriptive words: DYNO ('amazing'), ROASTER ('idiot').

4.2 Lexical variables

Our goal is to measure the rate at which people index their Scottishness (either consciously or subconsciously) through the use of distinctively Scottish words, and to find out whether this rate varies across different groups of users (Yes hashtag users vs. No hashtag users), or across different contexts (tweets which contain referendum-related hashtags vs. tweets that don't).

Were we to directly compare the frequencies of our Scottish terms across different sets of tweets, it would be difficult to untangle differences in the rate at which users are indexing the *referents* of those terms from differences in the rate at which they are indexing their Scottishness. For example, if people use the term MASEL ('myself') with a lower frequency in one context than in another, this could be because they are modulating their use of distinctively Scottish terms in response to the context, but it could also be because they are modulating the rate at which they talk about themselves. To avoid this confound, we instead compare the *conditional* probabilities with which Scottish terms are used, given that their referents are being indexed at all.

We therefore consider only those Scottish terms for which we can identify semantically equivalent Standard English variants. We require that each variant of a given variable indexes the same set of senses and can occur in the same set of contexts, so for example we do not include YOUS as a variant of **YOU**, since while Scottish YI and Standard English YOU can index both the singular and plural second person pronouns, YOUS is only used for the plural. We also did not include variants of **YES** and **NO** since their use could be influenced by campaign slogans (e.g., the hashtags *#VoteAye* and *#JustSayNaw*). Our variables are listed in Table 3.

5 Study 1: Scotland-specific vocabulary usage on either side of the debate

Do tweeters who use Yes hashtags use Scottish variants at a higher rate than tweeters who use No hashtags, either when using these hashtags, or in general?

5.1 Method

We assign users in the IT dataset to two groups, ***Yes*** and ***No***, based on the quantity $\frac{n_{u,yes}}{n_{u,yes}+n_{u,no}}$, where $n_{u,yes}$ is the number of tweets in which user u has used at least one of the Yes hashtags and none of the No hashtags in Table 1; and $n_{u,no}$ is the number of tweets in which u has used at least one No hashtag and none of the Yes hashtags. The ***Yes*** group consists of all users for whom this quantity is greater than or equal to 0.75, while the ***No*** group consists of all users for whom it is less than or equal to 0.25. Users for whom the value lies between 0.25 and 0.75 (as well as those for whom our dataset does not contain any tweets with Yes or No hashtags), are not assigned to either group. The ***Yes*** group

[6]While'bevy' is also used colloquially for 'beverage' in other parts of the UK, in Scotland it is more frequent and can additionally be used as a mass noun ("I had so much bevy I couldn't even carry it"), and as a verb ("I'd bevy with him every weekend").

Acronyms:	GTF MWI
Closed Class Words:	ABOOT AE AFF ATS DAE FAE HAE MASEL MASELF OAN OOR OOT TAE WAE WAN WI WIS YERSEL YI YIN YOUS
Contractions	CANNAE CANNI CANY CANA DEH DINI DINNY DIDNY DOESNY GONNAE GONY ISNY WANTY YER YIR
Discourse Markers:	ACH ANAW ANO AWRIGHT AWRITE AWRYT AYE EH NAE NAW OOFT YASS YASSS YASSSS YASSSSS YIP
Open Class Words:	AULD AWFY BAIRNS BAW BAWS BELTER BELTERS BEVY BOABY BOKE BRAW BURD BURDS CRABBIT DAFTY DAIN DEFOS DOON DUGS DYNO ECCIES FANNYS FEART FITBA FUD GAD GAWN GEES GID GRANDA GREETIN HAME HAW HING HINK HOOSE HOWLIN HUNNERS JIST LADDIE LASSIE LASSIES MANKY MAW MAWS MORRA MONGO PISH PISHED PISHING RAGIN ROASTER SARE SHITE SHITEY STEAMIN SUHIN WEANS WHITEY

Table 2: Scotland-specific vocabulary. Standard English equivalents of many words are shown in Table 3.

contains 4,513 users, while the ***No*** group contains 1,356 users, which is consistent with the general perception at the time that the Yes campaign was much more vocal than the No campaign. To test our hypothesis that the probability of choosing Scottish variants is, on average, greater for users in the ***Yes*** group than for users in the ***No*** group, we estimate the difference between the two groups in the average probability of choosing Scottish variants, and conduct a permutation test to approximate the distribution of this difference under the null hypothesis. We first test whether the ***Yes*** group are more likely than the ***No*** group to use Scottish variants in tweets which contain hashtags that indicate a stance on the referendum. Subsequently, we test whether the ***Yes*** group are more likely than the ***No*** group to use Scottish variants in general across all of their tweets.

5.1.1 Test statistic

Let U_g be the set of all users in group $g \in \{yes, no\}$ who have used at least one of the variables in Table 3. For a given user $u \in U_g$, let V be the set of all variables that u has used in at least one tweet. We estimate the probability of user u choosing a Scottish variant of variable $v \in V$ as $\hat{p}_{u,v} = \frac{n_{u,vscot}}{n_{u,v}}$, where $n_{u,vscot}$ is the token count of Scottish variants of v in user u's tweets, and $n_{u,v}$ is the token count of all variants of v in user u's tweets. Averaging across variables, we obtain $\hat{p}_u = \frac{1}{V}\sum_{v \in V} \hat{p}_{u,v}$. We then average across users to obtain the group mean, $\hat{p}_g = \frac{1}{U}\sum_{u \in U_g} \hat{p}_u$. Our test statistic is the difference between the two group means, $d = \hat{p}_{yes} - \hat{p}_{no}$.

5.1.2 Permutation test

We randomly shuffle users between the two groups (maintaining each group's original number of users), and re-compute the value of d using these permuted groups. We repeat this procedure 100,000 times in order to approximate the distribution of differences in group means that would be observable were the difference independent of the assignment of users to groups. The proportion of permuted differences which are greater than or equal to the observed difference between the original group means provides an approximate p-value.

	Tweets w/ Yes or No hashtags		All tweets	
Group	***Yes***	***No***	***Yes***	***No***
# Users	3776	1121	4352	1322
# Tweets	10,436	2411	173,171	80,736

Table 4: Number of users and tweets included per group in the two analyses in Study 1

5.2 Results

For a tweet to be included in the analysis, it must contain at least one of the variables in Table 3. Hence not all users contribute data to the test statistic, as some have not used any of the variables in their tweets. The number of tweets and users included in each analysis are shown in Table 4.

The results for the first analysis are shown in the left column of Table 5. The difference between the two groups in their average probability of choosing Scottish variants in tweets that contain polarised referendum hashtags is statistically significant ($p < 0.002$). Results for the second analysis are shown in the right column of Table 5. Once again, the difference between the two groups is statistically significant ($p < 0.001$).

5.3 Discussion

The results show that the ***Yes*** group do use Scottish variants at a significantly higher rate than the ***No*** group, both when using Yes or No hashtags, and in general. The stronger significance level for the 'All tweets' dataset is partly due to its larger size (see Table 4), which enables better estimates of the

Variable	Scottish variants (freq. per million words)	Standard English variants (freq. per million words)
ABOUT	ABOOT (50)	ABOUT (2562)
ALRIGHT	AWRIGHT (10), AWRITE (17), AWRYT (17)	ALRIGHT (77), ALL RIGHT (4)
BALL	BAW (11)	BALL (116)
BALLS	BAWS (17)	BALLS (47)
BIRD	BURD (35)	BIRD (78)
BIRDS	BURDS (31)	BIRDS (44)
DEFINITELY	DEFOS (27)	DEFINITIELY (217)
DIDNT	DIDNY (26)	DIDNT (563), DID NOT (31)
DO	DAE (61)	DO (2712)
DOESNT	DOESNY (18)	DOESNT (433), DOES NOT (33)
DOGS	DUGS (11)	DOGS (69)
DOING	DAIN (17)	DOING (590)
DONT	DEH (12), DINI (12), DINNY (62)	DONT (2880), DO NOT (92)
DOWN	DOON (49)	DOWN (786)
FOOTBALL	FITBA (13)	FOOTBALL (289)
FROM	FAE (77)	FROM (2485)
GIVES	GEES (14)	GIMME (5), GIVE ME (108), GIVE US (21), GIVES (75)
GOING	GAWN (15)	GOING (1884)
GOOD	GID (82)	GOOD (2602)
GRANDAD	GRANDA (7)	GRANDAD (19), GRANDFATHER (5), GRANDPA (9)
HAVE	HAE (9)	HAVE (4549)
HOME	HAME (22)	HOME (832)
HOUSE	HOOSE (20)	HOUSE (463)
I KNOW	ANO (42)	I KNOW (556)
ISNT	ISNY (16)	ISNT (342), IS NOT (151)
JUST	JIST (7)	JUST (5550)
MYSELF	MASEL (14), MASELF (15)	MYSELF (553)
OF	AE (75)	OF (9186)
OFF	AFF (82)	OFF (1567)
OLD	AULD (28)	OLD (526)
ON	OAN (38)	ON (7782)
ONE	WAN (33), YIN (28)	ONE(2537)
OUR	OOR (14)	OUR (790)
OUT	OOT (181)	OUT (3053)
PISSED	PISHED (19)	PISSED (66)
PISSING	PISHING (12)	PISSING (32)
SHIT	SHITE (428)	SHIT (764)
SHITTY	SHITEY (25)	SHITTY (52)
SOMETHING	SUHIN (17)	SOMETHING (614)
SORE	SARE (13)	SORE (140)
THATS	ATS (9)	THATS (1405)
THING	HING (11)	THING (749)
THINK	HINK (34)	THINK (1939)
TO	TAE (186)	TO (19996), TOO (1629)
TOMORROW	MORRA (27)	TOMORROW (1183)
WANT TO	WANTY (52)	WANNA (284), WANT TO (940)
WAS	WIS (33)	WAS (4197)
WITH	WI (85), WAE (116)	WITH (4774)
YOU	YI (26)	YOU (10891)
YOUR	YER (237), YIR (11)	YOUR (3094), YOURE (915), YOU ARE (342)
YOURSELF	YERSEL (11)	YOURSELF (193)

Table 3: Variables used in our studies, with each variant's frequency per million tokens in the GS dataset

	Tweets w/ Yes or No hashtags	All tweets
$\hat{p}_{yes}$	0.00766	0.01443
$\hat{p}_{no}$	0.00211	0.00734
d	0.00555	0.00709
p-value	0.00103	0.00001

Table 5: Results of the two analyses in Study 1

usage rates. While the rates are very low overall, the relative differences are large: the ***Yes*** group rate is more than three times the ***No*** group rate when we include only tweets with Yes or No hashtags, and approximately twice as big when we include all tweets. The higher rates in the 'All Tweets' dataset suggest that both groups of users chose Scottish variants less often when discussing the referendum than in their other tweets. However, the test we used does not provide a significance value for the difference in usage rates across the two datasets. To establish whether users do modulate their usage of Scottish variants when discussing the referendum, we will need a more careful paired design.

6 Study 2: Effects of topic and audience on Scotland-specific vocabulary usage

Do tweeters choose Scottish variants at a different rate when using referendum-related hashtags than in their other tweets?

6.1 Method

We need a statistic that corrects for the fact that some variables might have higher rates of Scottish variants than others. For example if users tend to produce Scottish variants of variable v_1 at a higher rate than for v_2, and use v_1 more in tweets that don't contain referendum-related hashtags, then it could appear that users are suppressing their Scottish usage in referendum-related tweets when in fact this is a lexical effect.

Let U be the set of all users who have used at least one of the variables in Table 3 in both a tweet that contains a referendum-related hashtag (i.e. a tweet that belongs to the IT dataset, referred to hereafter as an Indyref tweet) and in a tweet that does not contain a referendum-related hashtag (referred to hereafter as a Control tweet). For a given user $u \in U$, let V be the set of all variables that u has used in at least one Indyref tweet, and in at least one Control tweet. Let $\hat{p}_{I,v}$ for user u be the estimated probability that u chooses a Scottish variant of variable $v \in V$, conditioned on the fact that she is using variable v in an Indyref tweet. Analogously, let $\hat{p}_{C,v}$ be the estimated probability that u chooses a Scottish variant of variable v, conditioned on the fact that she is using variable v in a Control tweet. The difference in user u's probability of choosing a Scottish variant of variable v in an Indyref tweet and in a Control tweet is then $d_v = \hat{p}_{I,v} - \hat{p}_{C,v}$. Averaging across all variables, we define $d_u = \frac{1}{V} \sum_{v \in V} d_v$.

The null hypothesis is that on average, users are no more or less likely to choose Scottish variants in Indyref tweets than in Control tweets. Therefore, under the null hypothesis, the mean value of d_u across all users, $\bar{d}_u = \frac{1}{U} \sum_{u \in U} d_u$, would be zero. We perform a one-sample t-test to determine whether $\bar{d}_u$ is significantly different than zero.

We use this method to conduct two separate analyses. In the first analysis, our pool of Control tweets is the set of *all* tweets from the original filtered dataset that do not contain any of the hashtags in Table 1. In the second analysis, we limit our pool of Control tweets to those which do not contain any of the hashtags from Table 1, but *do* contain at least one other hashtag. This second analysis is designed to test whether the recent finding that US Twitter users are less likely to use regionally-specific words in tweets which contain hashtags (Pavalanathan and Eisenstein, 2015) applies to Scottish users as well.

6.2 Results

The number of tweets and users that were included in each analysis are shown in Table 6.

Results for the first analysis are shown in the left column of Table 7. The difference is statistically significant ($p < 0.01$), indicating that on average, individuals are less likely to choose Scottish variants when using referendum-related hashtags than in their other tweets. Results for the second analysis are shown in the right column of Table 7. In this case, the difference is not statistically significant.

6.3 Discussion

In light of (a) the apparent relationship between national identity and constitutional preference, (b) the history of Scots as the prestige language of a previously-independent Scotland, supplanted by English in large part due to the birth of the United Kingdom, and (c) the results of Study 1, which indicate that pro-independence users choose Scottish variants at a significantly higher rate than anti-

	All Controls	Controls w/ Hashtags
# Users	11,011	7429
# Indyref Tweets	41,924	35,241
# Control Tweets	693,815	195,145

Table 6: Number of users and tweets included in the two analyses in Study 2

	All Controls	Controls w/ Hashtags
$\bar{d}_u$	-0.0015	-0.0010
std error	0.0005	0.0006
t-statistic	-2.996	-1.758
p-value	0.0027	0.0788

Table 7: Results of the two analyses in Study 2

independence users—it may at first appear surprising that people are *less* likely to choose Scottish variants in tweets containing referendum-related hashtags than in their other tweets.

It is conceivable that ***Yes*** users increase their rate of Scottish variants in Indyref tweets whilst ***No*** users decrease it, such that their effects cancel out; but since ***Yes*** users are more prolific in the IT dataset, if anything we would expect this imbalance to make the effect even more positive. The fact that we see a significant negative effect in spite of the greater number of ***Yes*** tweets means we can be reasonably confident that even if ***Yes*** users aren't significantly reducing their usage of Scottish variants in Indyref tweets, they certainly aren't increasing it.

It is also worth noting that we did not exhaustively identify every hashtag that has been used in relation to the referendum, so inevitably there will be some tweets with referendum-related hashtags in the Control set (such as example tweet (3) in §2), and there may also be some non-referendum tweets in the Indyref set. However, if anything this would dilute any differences between the two lists, yet we still find an effect.

The fact that this effect does not reach significance when we remove Control tweets without hashtags suggests that the primary reason users are reducing their rate of Scottish variants in Indyref tweets is not because of the *topic* under discussion, but because the use of hashtags broadens the potential audience. This explanation accords with Pavalanathan and Eisenstein's (2015) finding that amongst Twitter users in the US, non-standard and regional variants are less likely to be used in tweets that target larger audiences. Of course, it is possible that topic has an effect as well, but the present study does not provide evidence for that conclusion.

7 Conclusion

We presented the first large-scale study of distinctively Scottish language use on social media, showing that this use includes a mixture of traditional Scots vocabulary, newer Scottish slang, and alternative spellings that reflect Scottish pronunciation. We also studied how users' language might reflect their political views and discourse. We showed that ***Yes*** users use Scottish variants at a higher rate than ***No*** users, whether discussing the independence referendum or not. But overall, users tend to decrease their use of Scottish variants when discussing the referendum. This result suggests that although ***Yes*** users generally express a stronger Scottish linguistic identity than ***No*** users, they are not choosing to express this identity strongly in political discourse aimed at a broad audience. Due to the very low rates of Scottish variants overall, our data set is too small to study differences between individual variables or even conclusively say whether there may be effects of both topic and audience size on the use of Scottish language. However, we hope to be able to answer these questions in future by collecting a more complete set of data for the particular users studied here.

8 Acknowledgements

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

References

Adam J. Aitken, James A.C. Stevenson, Sir William Alexander Craigie, and Margaret G. Dareau. 1990. *A Dictionary of the Older Scottish Tongue Vols. 1-7: From the Twelfth Century to the End of the Seventeenth*. MacMillan Publishing Company.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Lin Su Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social

media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130. Association for Computational Linguistics.

John Corbett, J. Derrick McClure, and Jane Stuart-Smith. 2003. A brief history of Scots. In John Corbett, J. Derrick McClure, and Jane Stuart-Smith, editors, *The Edinburgh Companion to Scots*, pages 1–16. Edinburgh, Edinburgh University Press.

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–106. Association for Computational Linguistics.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the International Conference on Machine Learning*, pages 1041–1048.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188.

Bruno Gonçalves and David Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PloS one*, 9(11):e112074.

William Grant and David D. Murison. 1931. *The Scottish National Dictionary*. Scottish National Dictionary Association.

William Grant. 1931. Phonetic description of Scottish language and dialects. In *The Scottish National Dictionary*, volume 1, pages 9–41. Online: `http://www.dsl.ac.uk/about-scots/history-of-scots/`.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2015. Understanding US regional linguistic variation with Twitter data analysis. *Computers, environment and urban systems*.

Taylor Jones. 2015. Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech*, 90(4):403–440.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18. Association for Computational Linguistics.

Billy Kay. 1988. *Scots: The Mither Tongue*. Grafton, first edition.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Caroline Mcafee. 1985. Nationalism and the Scots renaissance now. In Manfred Görlach, editor, *Focus on: Scotland (Varieties of English around the world, V.5)*, pages 7–16. Amsterdam/Philadelphia, John Benjamins Publishing Company.

Dong-Phuong Nguyen, RB Trieschnigg, and Leonie Cornips. 2015. Audience and the use of minority languages on Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 666–669.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Audience-modulated variation in online social media. *American Speech*, 90(2):187–213.

ScotCen. 2013. Should Scotland be an independent country? (combined responses of those who have and those who haven't decided yet) broken down by 'Moreno' national identity. Retrieved from: `http://whatscotlandthinks.org/`. Accessed: 2016-09-30.

Scottish Language Dictionaries. 2004. Dictionary of the Scots language. `http://www.dsl.ac.uk/`. Accessed: 2016-12-20.

Rachael Tatman. 2015. #go awn: Sociophonetic variation in variant spellings on Twitter. *Working Papers of the Linguistics Circle of the University of Victoria*, 25(2):97–108.